

To: UTC

Title: Supporting the Warsh orthography for Arabic script

From: Lorna Evans (SIL International)

Date: 29 April 2014

## Introduction

Recently my coworkers and I have been working on a font intended for use in West Africa. We wanted to support the Warsh (Warš) orthography (or reading tradition) for the Arabic script. Below is a description of three orthographic traditions. We list three potential solutions for Warsh orthographies and request feedback on which is the most appropriate Unicode solution.

### Different or Same Characters?

In the chart below we see how Warsh behavior contrasts with two other orthographic traditions - Hafs and Al-Duri. Hafs is the most common orthographic tradition for Modern Standard Arabic. Al-Duri is used in Sudan, Central Africa and some regions of Nigeria. Warsh is “the most widespread tradition in North and West Africa” (The Arabic Script in Africa, p 7). All of these traditions are used not only with Modern Standard Arabic but other languages in Africa as well.

### Hafs, Al-Duri and Warsh Compared

|                   | Hafs                | Al-Duri             | Warsh                                |
|-------------------|---------------------|---------------------|--------------------------------------|
| <b>feh</b>        | ف ففف<br>U+0641     | ب ببب<br>U+06A2     | ف ببب<br>U+06A1 U+06A2 U+06A2 U+06A1 |
| <b>qaf</b>        | ق ققق<br>U+0642     | ق ققق<br>U+0642     | ق ققو<br>U+066F U+06A7 U+06A7 U+066F |
| <b>noon</b>       | ن ننن<br>U+0646     | ن ننن<br>U+0646     | ن ننس<br>U+06BA U+0646 U+0646 U+06BA |
| <b>kaf</b>        | ك ككك<br>U+0643     | ك ككك<br>U+0643     | ك ككك<br>U+06A9                      |
| <b>imala e</b>    | -                   | ◌ِ<br>U+065C        | ◌ِ<br>U+065C                         |
| <b>“ii” final</b> | يِ<br>U+0650 U+0649 | يِ<br>U+0650 U+0649 | يِ<br>U+0650 U+06D2                  |

For the Hafs and Al-Duri orthographies it is a simple matter of choosing a different character. For example, the Hafs "feh" requires a nukta dot above and the Al-Duri "feh" requires a nukta dot below, but the four forms are consistent. The issue becomes more complex with Warsh. For some characters

(kaf, imala e and "ii" final) it is a simple matter of using another codepoint. In the Warsh orthography the isolate and final "feh", "qaf", and "noon" must be dotless. This is not currently supported in Unicode. The question is, should it be? Should someone using a Warsh orthography be able to use the same codepoint for isolate, initial, medial and final, or should they be required to change the codepoint depending on where in the word the character appears? In the chart above, we have used different codepoints for "feh", "qaf" and "noon" isolate and final forms than for the initial and medial forms.

Currently, many Microsoft fonts support the behavior we are looking for with U+06BA ARABIC LETTER NOON GHUNNA.

06BA ن نث (Arial)

06BA ن نث (Times New Roman)

06BA ن نث (Arabic Typesetting)

However, a Unicode document discussing U+06BA is located here: [L2/12-381](#). In response to L2/12-381, the Unicode Technical Committee decided to "Document in the standard and the nameslist that Noon Ghunna is dotless in all its contextual forms." (See [UTC minutes](#)). This apparently means the fonts supporting the Warsh noon behavior are no longer in Unicode compliance. Additionally, there do not appear to be any fonts which support the Warsh (dotless) forms for feh and qaf.

### Three Possible Solutions

There seem to be three options for supporting Warsh orthographies:

1. Use different codepoints depending on where in the word the character appears
  - a. **Pro:** Already able to use in Unicode
  - b. **Con:** This seems contrary to the whole spirit of Unicode. However, it is the current solution many are compelled to use due to a lack of other options.
  - c. **Con:** Searching becomes difficult. A search for "feh" would require searching for *either* U+06A2 *or* U+06A1.
  - d. **Con:** This pushes the data into a presentation encoding. Is there any difference between this approach and using presentation forms from U+FE00?
  - e. **Con:** If a word in a lexicon would be encoded differently from its form in a text, then there are serious technical problems with having different codes for initial and medial forms. In a dictionary, how is someone supposed to find a word based on an isolate character and then have to look somewhere else completely for the word itself?
2. Use one codepoint for "feh", another for "qaf" and another for "noon" and use character variants to support the isolate and final dotless forms and initial and medial dotted forms
  - a. **Pro:** No need to wait for adding characters to Unicode
  - b. **Pro:** Searching for one character is much easier
  - c. **Con:** Very few fonts and applications support character variants
  - d. If this is the preferred choice, which characters should the character variants be based upon, the dotless (U+06A1 ARABIC LETTER DOTLESS FEH, U+066F ARABIC LETTER DOTLESS QAF, U+06BA ARABIC LETTER NOON GHUNNA) or the dotted forms (U+06A2 ARABIC LETTER FEH WITH DOT MOVED BELOW, U+06A7 ARABIC LETTER QAF WITH DOT ABOVE, U+0646 ARABIC LETTER NOON)?

3. Add three new characters to Unicode to support the Warsh character requirements
  - a. **Pro:** Searching for one character is much easier
  - b. **Con:** It would be years before the characters are in Unicode and supported by fonts and applications

### Samples

Below are sample texts from a Warsh Qur-ān and a more standard Hafs Qur-ān:

| Warsh Qur-ān   | Hafs Qur-ān |
|--|-------------|
|  |             |
| Isolate and final noon (red), initial and medial qaf (green), initial and medial feh (turquoise) |             |
|  |             |
| Isolate feh (turquoise), medial noon (red)   |             |
|  |             |
| Isolate qaf (green), initial noon (red)  |             |
|  |             |
| Final qaf (green), final feh (turquoise)   |             |

### References

The Holy Qur-ān. English translation of the meanings and commentary. Revised & Edited By The Presidency of Islamic Researches, IFTA, Call and Guidance. Published 1989 by King Fahd Holy Qur-ān Printing Complex in Al-Madinah. Arabic text and English translation in parallel columns.

The Holy Qur-ān. Warsh Qur'an for Tijaniyya sect. Used in Cameroon, Nigeria and Niger.

Mumin, Meikal and Kees Versteegh (editors). 2014. The Arabic Script in Africa: studies in the use of a writing system. Studies in Semitic Languages and Linguistics. Vol. 71. Koninklijke Brill NV, Leiden, The Netherlands.

Pournader, Roozbeh. 2012. Initial and medial forms of Arabic Letter Noon Ghunna (L2/12-381 <http://www.unicode.org/L2/L2012/12381-dotless-noon.pdf>)

UTC. 2012. Approved Minutes of UTC Meeting 133. (L2/12-343 <http://www.unicode.org/L2/L2012/12343.htm>)